Authors' comments on anonymous Referee #1

We would like to thank the anonymous Referee #1 for the review and the very useful input to a possible revision of our manuscript. Please find below a point-to-point reply on the issues raised by Referee#1.

It is a bit difficult to identify the contribution made by this study as the findings presented largely confirm what is already known. Further, since they are based on single model simulations they are not robust.

We agree that simulations based on one single model do not provide robust climate change projections. To this end, a comparison to other model simulations has been included. However, this comparison was carried out in a qualitative way rather than a quantitative one.

Evaluation of the COSMO model is severely lacking due to the absence of perfect boundary experiments. As such it is impossible to properly validate the RCM (a task already made difficult due to the scarcity of observation, as the authors acknowledge), constrain the future changes and assess the added value of the RCM vis-à-vis the driving ESMs. The authors argue that validating historical simulations against observations is appropriate but need to provide a more convincing argument. There are substantial recent trends in precipitation over the Barents region and Scandinavia, which are due at least partly to internal variability, that
the historical simulations cannot be expected to reproduce.

As this point was also raised by Referee#2, and to provide a more convincing evaluation of the model, the ERAInterim driven simulation will be extended, and its evaluation included in the revised manuscript.

The study is also limited by the fact that it presents canonical results that are better suited to an ensemble approach. On the topic of future changes in precipitation, for example, the range of changes from a rather limited ensemble (10 members, maybe only 5 of which are independent) of high resolution downscalings from CORDEX is -2% - +33% in winter and -12% - +23% in summer over western Norway. While the COSMO simulations presented here certainly fall within these bounds they do not add any new information.

We agree that the COSMO simulations do not provide any new values to the range of possible changes. But, as they fall within the bounds, they are enhancing the likelihood of changes within these bounds.

With respect to sea-ice the present day bias in the EC-Earth
simulation renders the changes indicated by the future simulation highly suspect as they are so clearly linked to this field.

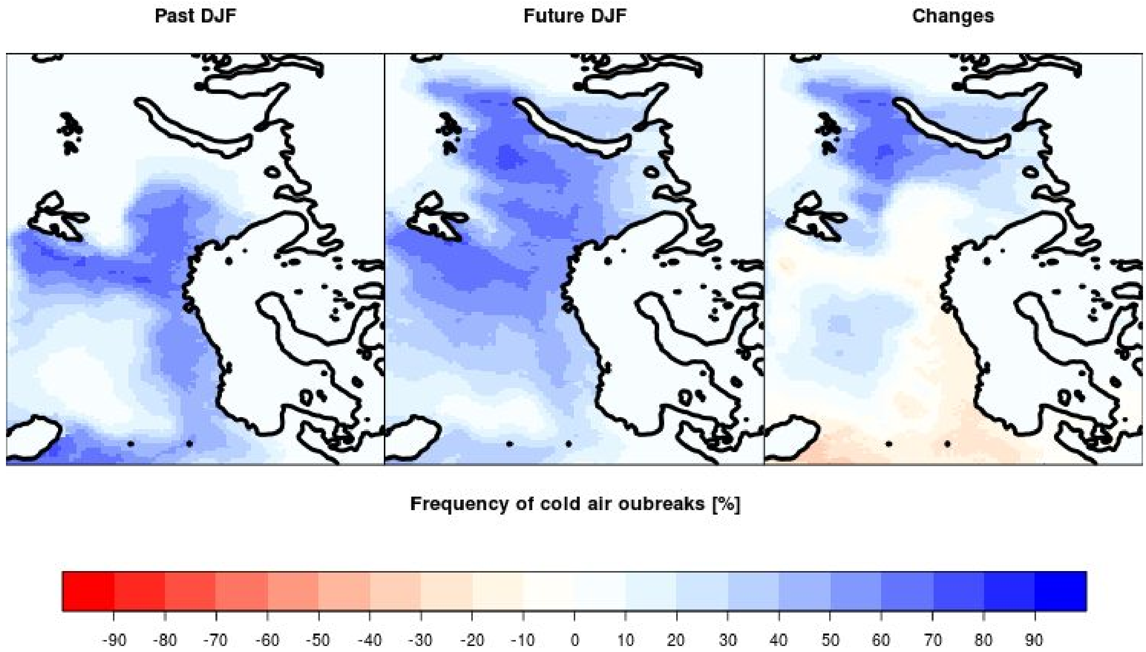
As pointed out by both referees, the present day sea-ice bias in the EC-Earth driven runs is very large and make the projected changes highly suspect. Thus, we will skip the EC-Earth driven runs in the revised manuscript.

This reviewer also struggled to find the added value described in the last sentence of the abstract. NorESM is clearly an outlier but since it is not downscaled with COSMO so it cannot exhibit any added value with respect to this driving model. Once it is removed the RCMs and ESMs largely agree on cloud cover decline in both sign and magnitude. In the text the authors themselves state that COSMO exhibits a similar reduction in cloud cover as the driving MPI-ESM. So where has the added value come from exactly? Koenigk et al. (2015) showed that there was little added value obtained by the regional models for canonical variables
and processes due to the strong dependence on the large-scale features of the driving ESMs. This study does little to change that.

We argue that the RCM ensemble (Koenigk et al. ensemble + the DMI-HIRHAM + COSMO run) shows an added value compared to the driving ESMs, as the RCM ensemble shows much less spread than the ESM ensemble. In the ESM ensemble, not only the NorESM, but also the CanESM2 shows a positive trend. For both ESMs, the RCA4 RCM shows a negative trend, in agreement with the other RCM projections. Additionally to the findings of Koenigk et al. (that the RCA4 cloud cover is independent of the GCM) we find that the projected cloud cover trend is also rather independent of the RCM used (RCA4, HIRHAM, COSMO). However, we can see the point that calling the smaller spread an added value may not be considered accurate without evaluating the simulated cloud cover. As estimates of the true winter cloud cover in this area are very uncertain, a proper evaluation will not be possible. Thus, we will not phrase it as an added value in the revised manuscript but point out the fact that the projected RCM trends seem to be largely independent of the models used.

Single model experiments suffer from a lack of robustness. Their best application is to test model performance and the simulation of relevant processes. Might it not be more useful then, given the limitations of a single model experiment design, to investigate processes where the RCM might be expected to show improvement, such as surface fluxes and how they relate to the changes in cloudiness for example? Or assess cold air outbreaks, confirm that COSMO can reproduce these events and then say something about future changes. Or even better actually explore the dependencies between variables as described in the introduction?

We've started calculating cold air outbreaks (based on the MCAO index by Fletcher et al. 2016, http://dx.doi.org/10.1175/JCLI-D-15-0268.1) in the simulations and an analysis will be included in the reviewed manuscript. First results for past and future cold air outbreak frequencies show (see figure below) arge increases where sea-ice is retreating, some decreases in the southern parts and appearance of cold air outbreaks in the Kara sea as well as north of Svalbard in the future winter projections. The reviewed manuscript will include an analysis of cold air outbreaks in the ERAInterim driven COSMO run to see if the model is able to reproduce the current events and recent trends as well as future projections.



Despite its shortcomings this work has potential. However, the authors must make an effort to provide something more than simple an addition to the Koenigk et al. (2015) ensemble. As it stands this is the only contribution made, as the added value claimed is simply not supported by the results. Major revisions are needed with new and perhaps a restructured analyses. While I would like to see an evaluation simulation I do not believe it is necessary if the authors can expand the scope of their evaluation (CRU is too limited and questionable in this region). Given the scope of the revisions required I am supplying only major comments at this time. Minor comments will follow once a fully revised manuscript is submitted.

Major Comments: 1. In the introduction the authors propose a number of potentially fruitful avenues of investigation but then fail to follow up on them. The possible exception might be the cloud cover issues but this proved not to be very informative (see comment X). The paper begins by stating that the dependencies between e.g., cloud cover, precipitation type and wind speeds are under-investigated. This is a promising line of investigation. However, it is not pursued as the authors then go on to present and discuss the changes in said fields in isolation. Some deeper investigation (such as covariation of winds and convective precipitation, temperature and pressure, temperature and precipitation, etc.) would greatly improve the quality of the manuscript. This is also where analysis of a single model set of experiments is most useful.

Also some dependencies are shown, we agree that our analyzing of the variables is still too focussed on the single variables. In the revised manuscript, we will include more fields to fill *these gaps (e.g. winds on different levels, cold air outbreaks, fluxes and pressure) and put more emphasize on the interactions between the single variables.*

2. The introduction goes on to mention placing the current study within the scope of empirical statistical approaches such as those employed by Benestad et al. (2016). This also was not explored further. Setting the COSMO simulations within the scope of the ESD work could provide an interesting space to explore the strengths and weaknesses of both approaches. Are there things we learn from the COSMO experiments that we do not learn from the Benestad et al. (2016) downscalings? Obviously the benefit of the ESD approach is that it is statistically robust. But is it physically plausible?

We consider to elaborate more on the physical and statistical robustness and limitations of both approaches. However, depending on the changes of the other parts of the manuscript we may also skip this issue completely to keep the manuscript more focussed.

3. On lines 52-54 the mini-ensemble from Koenigk et al. (2015) is introduced. If the authors intend to "sell" this study as an additional ensemble member they need to be clear on this issue. The additional ensemble members need to be introduced and the reasons for, and scope of, their inclusion described much earlier in the text (Introduction and Methods sections). It is quite confusing when in the discussion of results to suddenly shift from the single model perspective to the ensemble perspective. Since one of the stated aims is to increase the ensemble size of available CMIP5 downscalings, I would suggest the entire analysis (including Pr, T and sea-ice) be performed on the existing ensemble (Koenigk et al. (2015) + COSMO). In this way the analysis will at least be consistent throughout although in practice the simulations likely have quite different skillful scales as the Arctic-CORDEX runs are at .44 and the COSMO runs at .22 resolutions. Some discussion on the appropriateness of the comparison is required.

The other RCM runs have mainly been introduced to assess the robustness issue with the single model experiment. We will consider the idea of performing the whole analysis (or at least the parts where the data is available) on the whole ensemble and make this more clear in the reviewed manuscript and introduce the ensemble much earlier.

4. As stated previously, the lack of robustness is an issue. That the COSMO simulations are similar to other simulations is not terribly illuminating. Since robustness cannot be fully evaluated here I suggest addressing the following questions: Are the changes presented outside the confidence intervals of the historical simulation? Are they statistically significant (preferably with a non-parametric Monte Carlo-type approach)?

We will consider tests of statistical significance in the revised manuscript, taking the historical variability into account. Robustness will also be partly covered by changing from a single model framework to an ensemble approach in the revised manuscript.

5. The lack of realism in the present EC-Earth sea-ice renders the changes described highly suspect. It is also unclear why only the rcp4.5 simulation is included when rcp2.6 and rcp8.5 are performed for MPI-ESM. The entire experiment design has the feeling of incompleteness. It comes across as though this was an exercise in downscaling rather than to address specific scientific questions. As result it is difficult to discern what it obtained through the inclusion of the EC-Earth. The authors claim on L261-264 that the EC-Earth runs provide important insights but decline to provide any details as to what these are. Something much more specific such as the demonstrated effect of the sea-ice bias on the future changes is needed else just take out the EC-Earth Simulation.

We will take out the EC-Earth simulation.

6. The evaluation of the COSMO simulations over the historical period is also puzzling. Why is CRU, a land-based, interpolated, gridded dataset, used? Given that the domain of interest is mostly over water this decision is baffling. The precipitation from interpolated datasets with sparse observations is highly suspect. The temperature might be more reliable but again will suffer due to station sparseness. Inclusion of the MERRA reanalysis is welcomed but the way it is included is not terribly informative as it is compared against CRU with CRU held up as ground truth. I would recommend comparing the COSMO simulations directly against more reanalysis products (ERA-Interim, MERRRA, Arctic System Reanalysis (http://polarmet.osu.edu/ASR/)) and provide considerably more detail and explanation as to why CRU is in any way a reliable ground truth for the Barents region. My argument would be that none of these products are sufficient in and of themselves but that taken collectively one can cobble together a reasonable picture of model performance. As mentioned previously, just looking at modeled temperature and precipitation biases is not so informative and has been done before. How about looking at processes that COSMO, or any model, has to get right in order to produce a plausible reproduction of climate over the Barents region?

We will include an ERAInterim driven COSMO run and evaluate this in more details. Clearly, an evaluation based on CRU only is not appropriate in the region. This was also the argument to introduce MERRA: It was meant as an alternative "truth" in the region as CRU can not be considered the ground truth. However, we can see now that the limitation of MERRA to land points and the comparison of CRU to MERRA is more puzzling than clarifying and this will be changed in the revised manuscript.

7. The one aspect where the authors claim to demonstrate added value of the COSMO simulations is cloud cover. However this proves not to be the case. COSMO simply reproduces the cloud cover trend shown by the driving MPI-ESM (Figure 10). If the argument instead is that collectively (i.e., the Koenigk et al. ensemble + COSMO) these experiments exhibit added value then this should be demonstrated quantitatively. Just claiming they add value because they agree is not especially convincing. Especially when they largely reproduce the ESM trends with one outstanding exception

(NorESM).

Yes, the argument was that the RCM ensemble (Koenigk et al. ensemble + the DMI-HIRHAM + COSMO run) shows an added value compared to the driving ESMs, as the RCM ensemble shows much less spread than the ESM ensemble. However, we can see the point that calling the smaller spread an added value may not be considered accurate without evaluating the simulated cloud cover. Thus, we will not phrase it as an added value in the revised manuscript but point out the fact that the projected RCM trends seem to be largely independent of the models used. Unfortunately, we do not see a way to demonstrate the added value quantitatively due to the lack of reliable estimates of winter cloud cover in the Barents region.

8. The partitioning of clouds and convective/non-convective precipitation is potentially interesting and again would benefit greatly from a more detailed analysis focused on the relevant processes (e.g., cold air outbreaks, winds, fluxes). Given the conclusion that the increase in convective clouds goes along with a decrease in total and low level cloud the exclusion of the plot for COSMO showing changes in the lower levels is unwarranted (L215-217).

We agree that from a scientific point of view, changes in wind direction (also on different model or pressure levels) may be more relevant than changes in wind gusts only. Thus, wind changes on different height-levels will be included. We have also started calculating cold air outbreaks and how they are related to the wind and sea-ice changes. Projections and an analysis will be included in the reviewed manuscript. For the COSMO, cloud changes at the lower levels will be added to the manuscript.